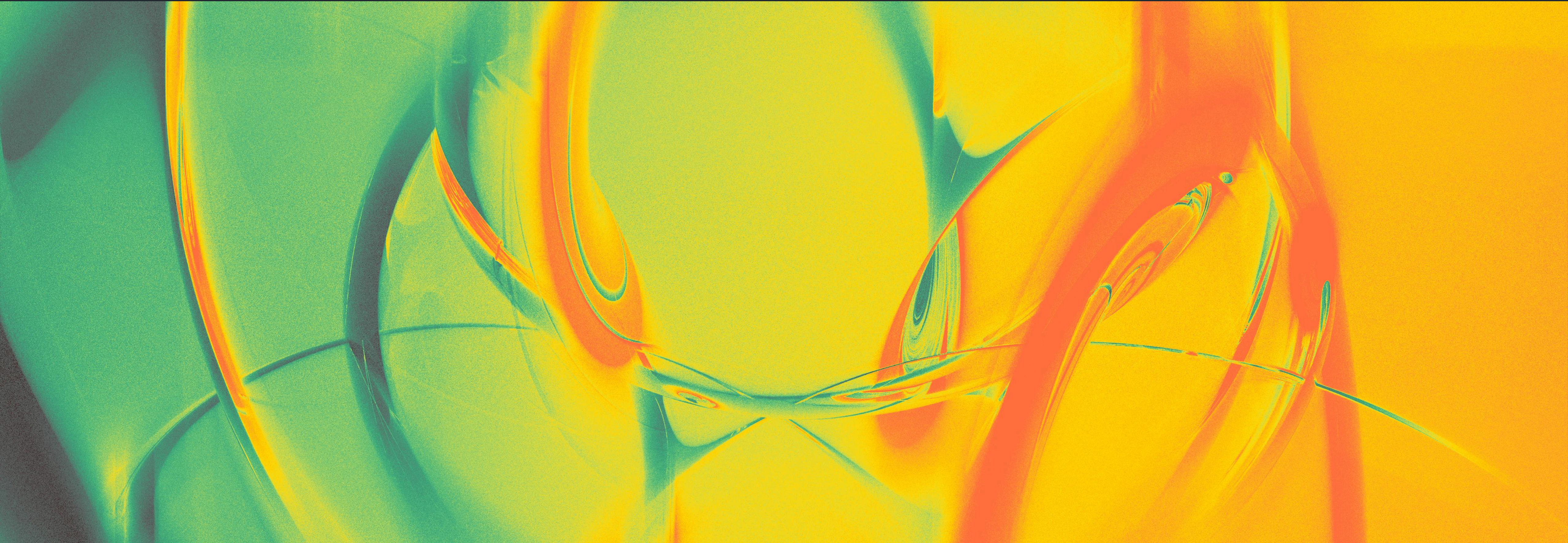


AI Evals in Vertical Industries

Guest lecture for “AI Evals For Engineers & PMs” course

Dr Christopher Lovejoy, MD
Head of Clinical AI

2025-08-07
Anterior



OUTLINE

- Why is it harder to do evals for specialized verticals?
- Empowering **domain experts** to look at your data
- Principles for building an effective domain expert **review dashboard**
- Using **review data** to power your system
- How to hire the right **domain experts**

What defines a “specialized vertical”?

What defines a “specialized vertical”?

Could a lay person develop “top tier” judgment
after working on this for 2 months?

**Why is it harder to do evals for
specialized verticals?**

Why is it harder to do evals for specialized verticals?

- 1 . It's harder to define quality / correctness

Q: Does the patient have a prior brain MRI suspicious for Multiple Sclerosis?

 Florence:

The medical record shows a brain MRI from 17/5/2022 that demonstrates 'multiple foci of T2/FLAIR hyperintensity in the infratentorial, juxtacortical and periventricular white matter' which is noted to be 'consistent with multiple sclerosis.' This confirms prior brain MRI findings suspicious for MS.

Q: Does the patient have a prior brain MRI suspicious for Multiple Sclerosis?

 Florence:

The medical record shows a brain MRI from 17/5/2022 that demonstrates 'multiple foci of T2/FLAIR hyperintensity in the infratentorial, juxtacortical and periventricular white matter' which is noted to be 'consistent with multiple sclerosis.' This confirms prior brain MRI findings suspicious for MS.

Why is it harder to do evals for specialized verticals?

- 1 . It's harder to define quality / correctness
- 2 . It's harder to define failure modes

Example failure modes: clinical reasoning

Failure Mode	Description
<i>Retrospective reasoning</i>	Incorrectly using evidence that became available after the decision point to justify whether the patient should have had the procedure in the first place.
<i>Under-inference</i>	Not making valid inference, e.g saying something ‘needs to be explicitly stated’ when it could be reasonably inferred
<i>Over-inference</i>	Drawing conclusions that go beyond what the evidence supports. Making assumptions without sufficient basis.
<i>Misunderstanding chronology</i>	Incorrect interpretation or application of the sequence of events (reading events in wrong order)
...	...

Why is it harder to do evals for specialized verticals?

- 1 . It's harder to define quality / correctness
- 2 . It's harder to define failure modes
- 3 . It's harder to write prompts
 - a . Prompts in your main pipelines
 - b . LLM-as-judge prompts to evaluate your pipelines

Why is it harder to do evals for specialized verticals?

Error analysis

- 1 . It's harder to define quality / correctness
- 2 . It's harder to define failure modes
- 3 . It's harder to write prompts
 - a . Prompts in your main pipelines
 - b . LLM-as-judge prompts to evaluate your pipelines

Why is it harder to do evals for specialized verticals?

Error analysis

- 1 . [It's harder to define quality / correctness]
- 2 . [It's harder to define failure modes]
- 3 . It's harder to write prompts
 - a . Prompts in your main pipelines
 - b . [LLM-as-judge prompts to evaluate your pipelines]

Implementing Automated Evaluators

Why is it harder to do evals for specialized verticals?

Error analysis

- 1 . [It's harder to define quality / correctness]
- 2 . [It's harder to define failure modes]
- 3 . It's harder to write prompts
 - a . [Prompts in your main pipelines]
 - b . [LLM-as-judge prompts to evaluate your pipelines]

Improvement

Implementing Automated Evaluators

**The solution: bring domain experts
into the loop**

**The solution: bring domain experts
into the loop**

But where? and how?

“look at your data”

“look at your data”

The medical record shows a brain MRI from 17/5/2022 that demonstrates 'multiple foci of T2/FLAIR hyperintensity in the infratentorial, juxtacortical and periventricular white matter' which is noted to be 'consistent with multiple sclerosis.' This confirms prior brain MRI findings suspicious for MS.

“look at your data”

The medical record shows a brain MRI from 17/5/2022 that demonstrates 'multiple foci of T2/FLAIR hyperintensity in the infratentorial, juxtacortical and periventricular white matter' which is noted to be 'consistent with multiple sclerosis.' This confirms prior brain MRI findings suspicious for MS.

“but what does this mean?”



“look at your data” —————→

“empower
domain experts
to look at (and
translate) your
data”

The medical record shows a brain MRI from 17/5/2022 that demonstrates 'multiple foci of T2/FLAIR hyperintensity in the infratentorial, juxtacortical and periventricular white matter' which is noted to be 'consistent with multiple sclerosis.' This confirms prior brain MRI findings suspicious for MS.

“but what does
this mean?”



Empowering domain experts
to look at your data

Example 1: Raw traces

The screenshot displays the Langfuse interface for a trace. The top navigation bar shows the environment 'PROD-EU', the project 'langfuse-dev', and the team 'Team'. The breadcrumb trail indicates the current view is 'docs-examples / Traces'. The main header shows the trace ID 'main: ba27e7b1-e23e-4f50-87de-420cf038190f' and includes search, star, and share icons. On the left sidebar, there is a search bar and a list of steps in the trace. The main panel on the right provides a detailed view of the selected 'main' step.

Trace Details:

- Trace ID:** main: ba27e7b1-e23e-4f50-87de-420cf038190f
- Environment:** default
- Latency:** 1.24s
- Total Cost:** \$0.000763
- Token Usage:** 650 → 113 (Σ 763)

Steps in Trace:

Step Name	Latency	Cost
main	1.24s	\$0.000763
classify_feedback	1.02s	\$0.000194
OpenAI-generation	1.02s	\$0.000194
classify_feedback	1.23s	\$0.00019
OpenAI-generation	1.22s	\$0.00019
classify_feedback	1.01s	\$0.000187
OpenAI-generation	1.00s	\$0.000187
classify_feedback	0.95s	\$0.000192
OpenAI-generation	0.94s	\$0.000192

main Step Details:

Input:

```
{  
  "args": [  
    0: [  
      0: "The chat bot on your website does not work."  
      1: "Your customer service is exceptional!"  
      2: "Could you add more features to your app?"  
      3: "I have a question about my recent order."  
    ]  
  ],  
  "kwargs": {}  
}
```

Output:

```
[  
  0: {  
    "feedback": "The chat bot on your website does not work."  
    "classification": [  
      0: "BUG"  
    ]  
    "relevance_score": 0.9  
  },  
  1: {  
    "feedback": "Could you add more features to your app?"  
    "classification": [  
      0: "SUGGESTION"  
    ]  
    "relevance_score": 0.8  
  },  
  2: {  
    "feedback": "Your customer service is exceptional!"  
    "classification": [  
      0: "PRAISE"  
    ]  
    "relevance_score": 0.9  
  },  
]
```


Example 2: A custom view within the product

Avila

1

Tuesday, June 3, 2025

Patient Details

NAME:
Karlie

PHONE NUMBER:
+15559991236

Blueprint

ONBOARDING

Messages

Show Rationale

Approval Needed

CONFIDENCE: 70%

Tuesday, May 27

RATIONALE

Initial message

Today

PT

patient 6/3/2025, 10:24:21 AM

Yes I got it from the clinic when they sent me home

ava 6/3/2025, 10:25:24 AM

CONFIDENCE: 100%

Great, you're ready to start your treatment. Last thing - can you tell me what time it is where you are, right now? We won't use this information in any way other than to help keep you on your treatment schedule.

RATIONALE

Patient confirmed they have their medication from the clinic. Following the onboarding blueprint Message 0.7 to determine their local time/timezone so we can schedule future messages correctly. This is required before proceeding to the Step 1 + Step 2 treatment module.

PT

patient 6/3/2025, 10:26:15 AM

Sure it's 7:26

Patient Details

Timezone Offset updated to -420

Treatment phase updated to step1step2

Current Messages

a

Thanks, let's get started. Have you already taken the first pill?

CONFIDENCE: 100%

SEND NOW

RATIONALE

Following the Step 1 + Step 2 blueprint Message 1.1. The patient has confirmed they have their medication and I've determined their timezone offset (Pacific Time, -420 minutes). Now transitioning to the Step 1 + Step 2 treatment phase by asking if they've taken the first pill. This is the first question in the treatment sequence to determine their progress with the medication regimen.

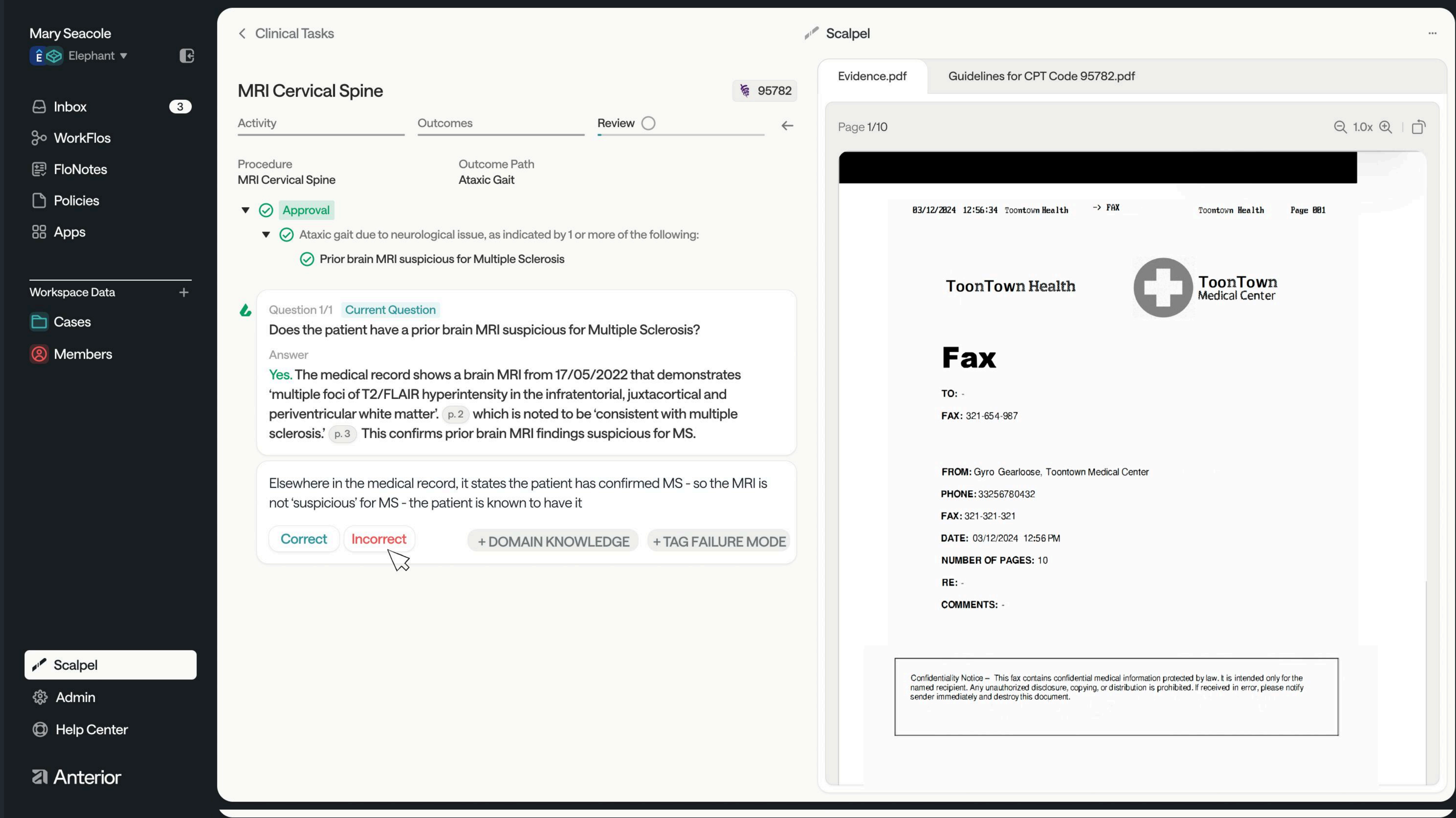
a

Just checking in. Have you decided when to take Step 1?

CONFIDENCE: 100%

SEND NOW

Example 3: A custom review dashboard



Principles for building an effective domain expert review dashboard

Optimise for 3 things:

- 1 . High quality reviews
- 2 . Minimise time per review
- 3 . Generate actionable data

Optimise for 3 things:

- 1 . High quality reviews
- 2 . Minimise time per review
- 3 . Generate actionable data



Principle 1: Optimise for clearly surfacing all required context

Principle 2: Optimise the review flow sequence

Principle 3: Design reviews that give the data you need

**Principle 1: Optimise for clearly
surfacing all required context**

Mary Seacole
Elephant

Inbox

WorkFlos

FloNotes

Policies

Apps

Workspace Data

Cases

Members

Scalpel

Admin

Help Center

Anterior

Clinical Tasks

MRI Cervical Spine

95782

Activity

Outcomes

Review

Procedure

MRI Cervical Spine

Outcome Path

Ataxic Gait

Approval

Ataxic gait due to neurological issue, as indicated by 1 or more of the following:

Prior brain MRI suspicious for Multiple Sclerosis



Question 1/1 Current Question

Does the patient have a prior brain MRI suspicious for Multiple Sclerosis?

Answer

Yes. The medical record shows a brain MRI from 17/05/2022 that demonstrates 'multiple foci of T2/FLAIR hyperintensity in the infratentorial, juxtacortical and periventricular white matter'. p.2 which is noted to be 'consistent with multiple sclerosis.' p.3 This confirms prior brain MRI findings suspicious for MS.

Elsewhere in the medical record, it states the patient has confirmed MS - so the MRI is not 'suspicious' for MS - the patient is known to have it

Correct

Incorrect

+ DOMAIN KNOWLEDGE

+ TAG FAILURE MODE

Scalpel

Evidence.pdf

Guidelines for CPT Code 95782.pdf

Page 1/10

1.0x

03/12/2024 12:56:34 Toontown Health -> FAX Toontown Health Page 001

ToonTown Health



ToonTown
Medical Center

Fax

TO: -

FAX: 321-654-987

FROM: Gyro Gearloose, Toontown Medical Center

PHONE: 33256780432

FAX: 321-321-321

DATE: 03/12/2024 12:56 PM

NUMBER OF PAGES: 10

RE: -

COMMENTS: -

Confidentiality Notice - This fax contains confidential medical information protected by law. It is intended only for the named recipient. Any unauthorized disclosure, copying, or distribution is prohibited. If received in error, please notify sender immediately and destroy this document.

Principle 2: Optimise the review flow sequence

Review case summary



**Understand current
decision point**

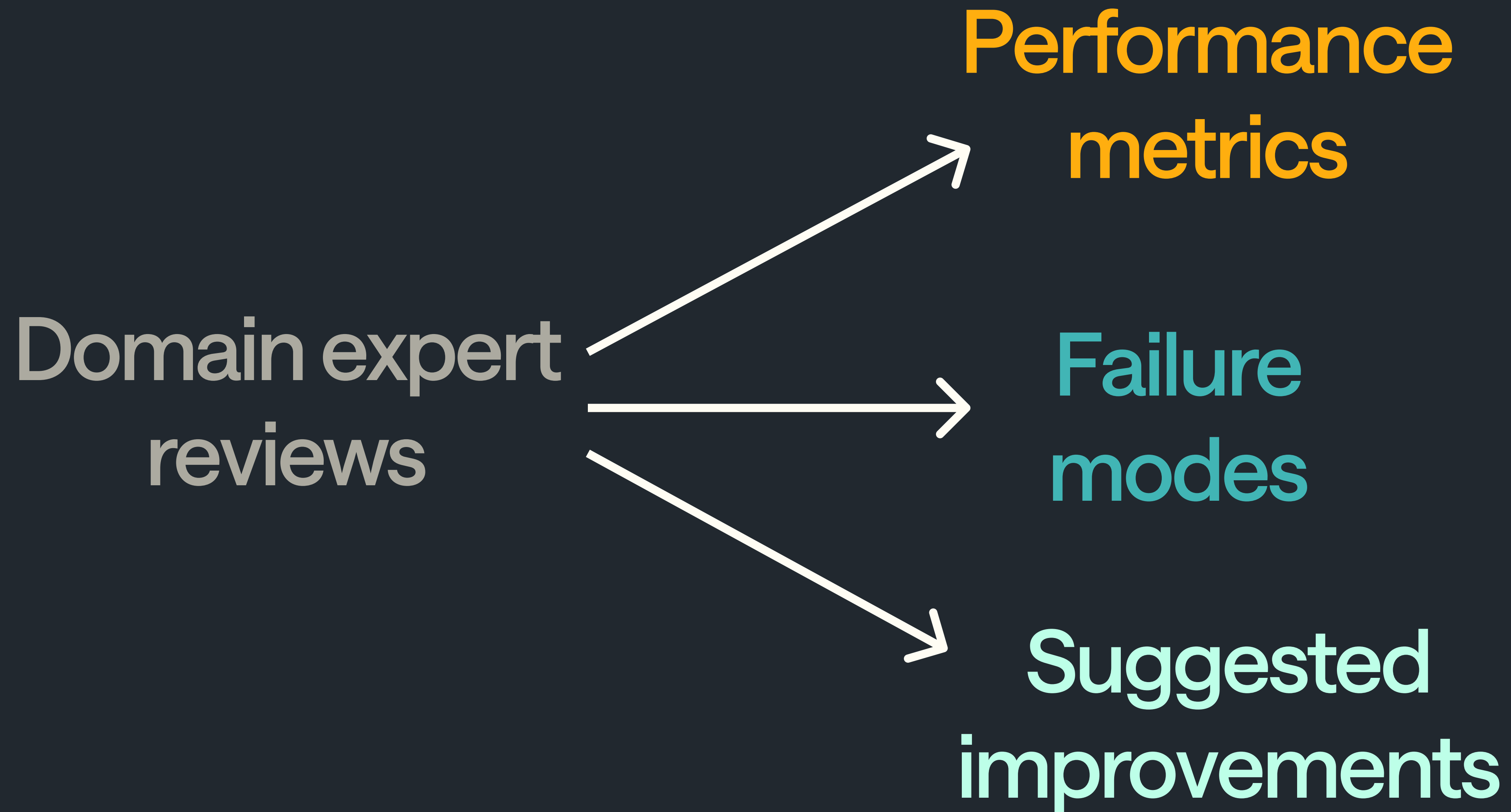


**Review the relevant
medical evidence**



**Appraise the AI
output**

**Principle 3: Design reviews that
give the data you need**



Mary Seacole
Elephant

Inbox

WorkFlos

FloNotes

Policies

Apps

Workspace Data

Cases

Members

Scalpel

Admin

Help Center

Anterior

Clinical Tasks

MRI Cervical Spine

95782

Activity

Outcomes

Review

Procedure

MRI Cervical Spine

Outcome Path

Ataxic Gait

Approval

Ataxic gait due to neurological issue, as indicated by 1 or more of the following:

Prior brain MRI suspicious for Multiple Sclerosis



Question 1/1 Current Question

Does the patient have a prior brain MRI suspicious for Multiple Sclerosis?

Answer

Yes. The medical record shows a brain MRI from 17/05/2022 that demonstrates 'multiple foci of T2/FLAIR hyperintensity in the infratentorial, juxtacortical and periventricular white matter'. p.2 which is noted to be 'consistent with multiple sclerosis.' p.3 This confirms prior brain MRI findings suspicious for MS.

Elsewhere in the medical record, it states the patient has confirmed MS - so the MRI is not 'suspicious' for MS - the patient is known to have it

Correct

Incorrect

+ DOMAIN KNOWLEDGE

+ TAG FAILURE MODE

Scalpel

Evidence.pdf

Guidelines for CPT Code 95782.pdf

Page 1/10

1.0x

03/12/2024 12:56:34 Toontown Health -> FAX Toontown Health Page 001

ToonTown Health



ToonTown Medical Center

Fax

TO: -

FAX: 321-654-987

FROM: Gyro Gearloose, Toontown Medical Center

PHONE: 33256780432

FAX: 321-321-321

DATE: 03/12/2024 12:56 PM

NUMBER OF PAGES: 10

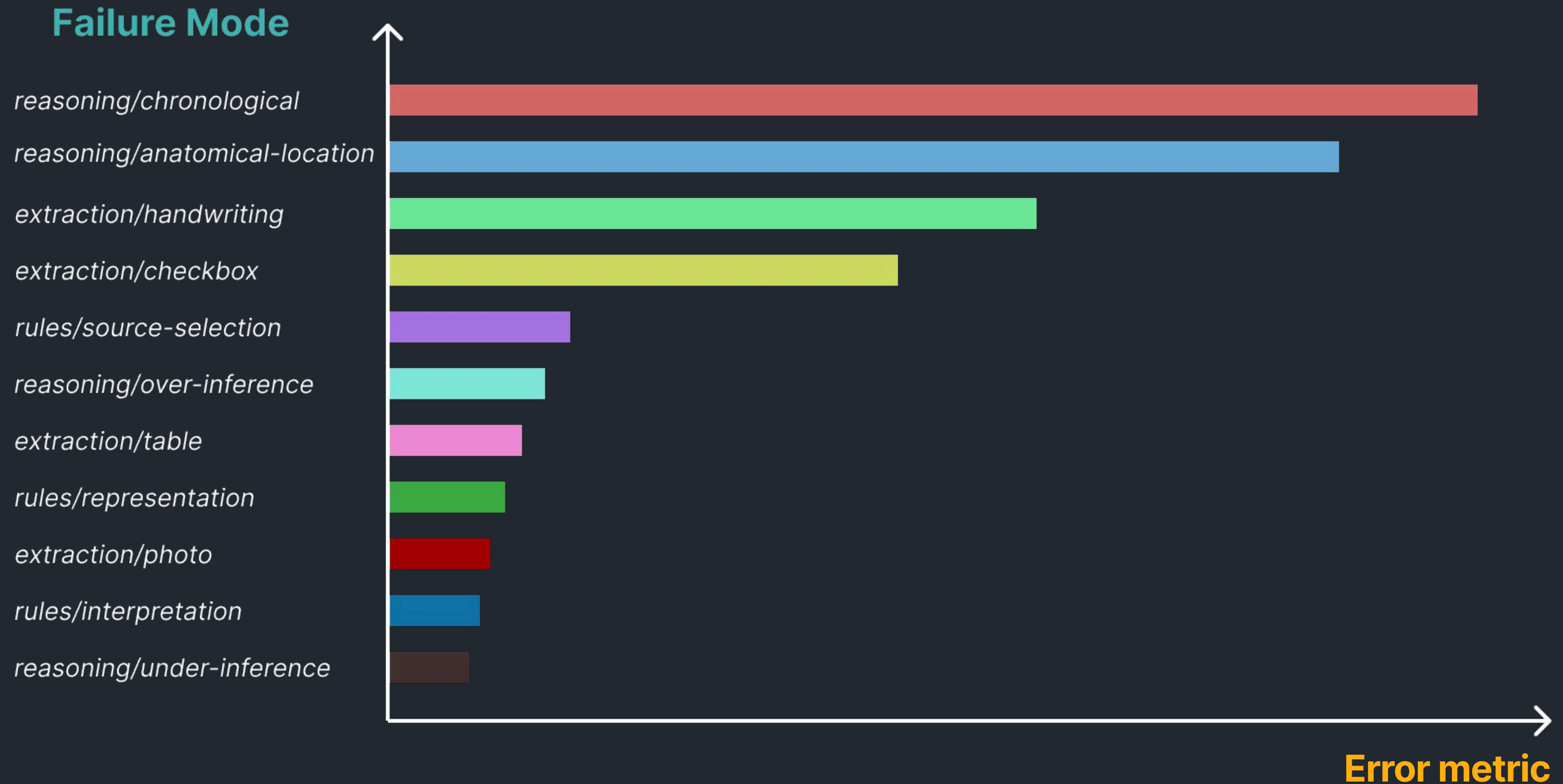
RE: -

COMMENTS: -

Confidentiality Notice - This fax contains confidential medical information protected by law. It is intended only for the named recipient. Any unauthorized disclosure, copying, or distribution is prohibited. If received in error, please notify sender immediately and destroy this document.

Using review data to
power your system

(1) Use metrics and failure modes from production data to prioritise work



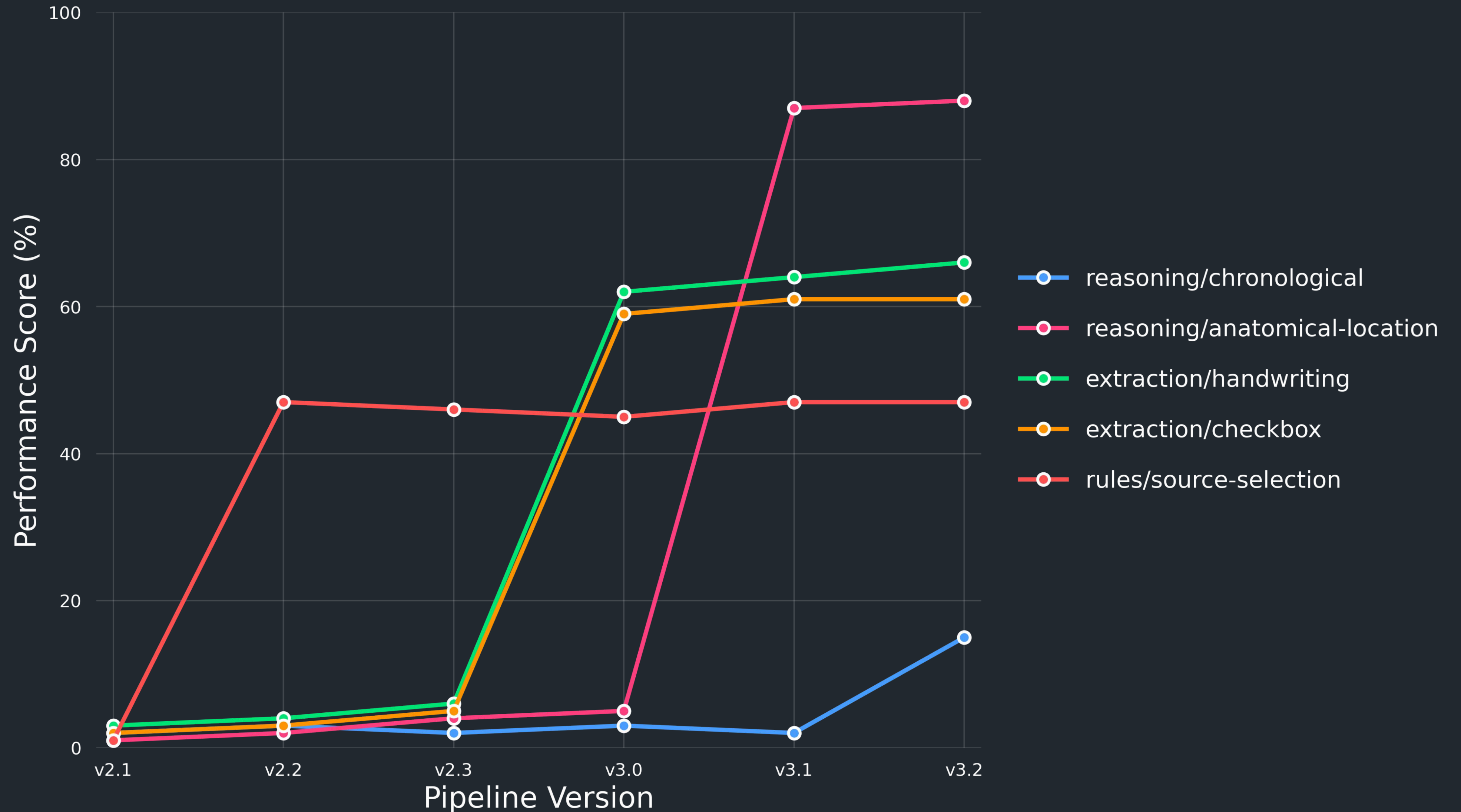
(2) Use failure mode datasets to test the impact of changes

Clinical Reasoning >> Over-inference

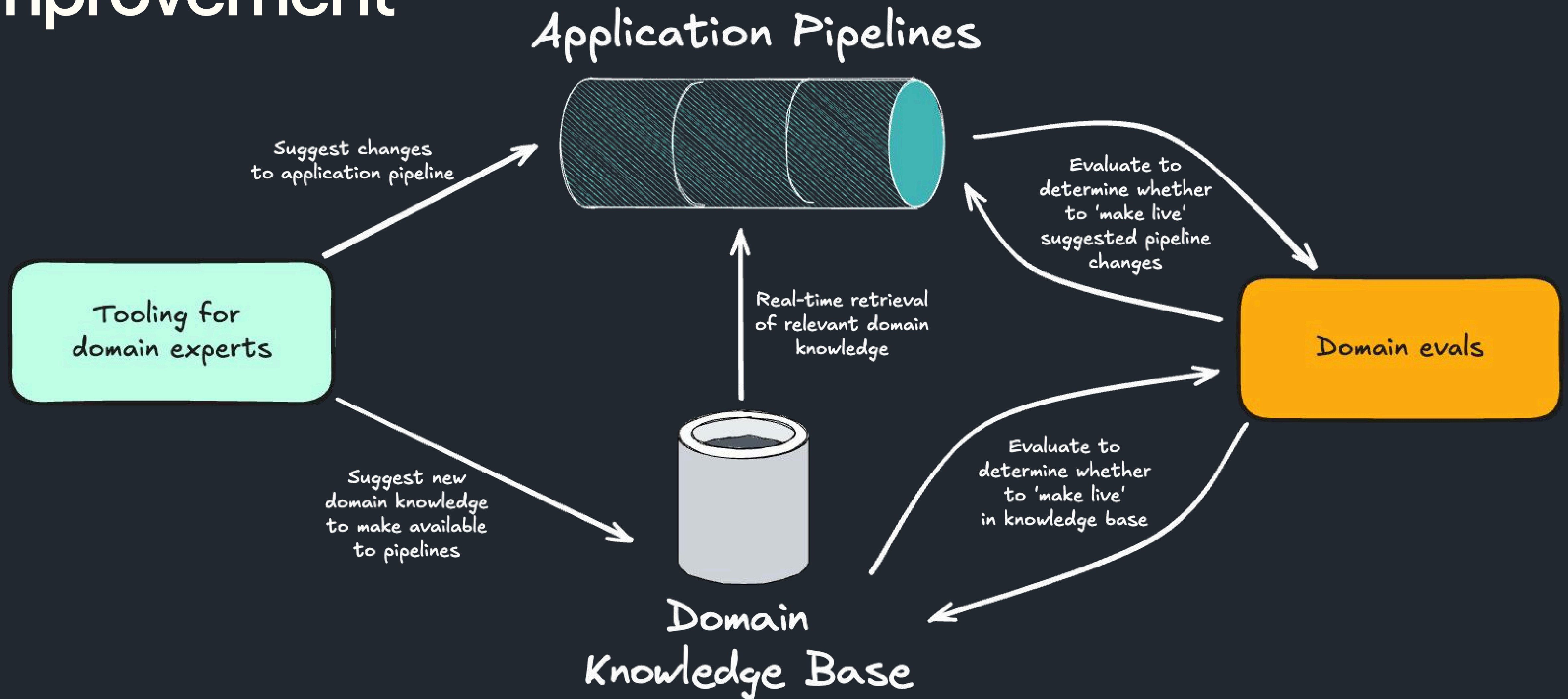
Medical Record Extraction >> Table Extraction

Rules Interpretation >> Logic representation

(2) Use failure mode datasets to test the impact of changes



(3) Create mechanisms for automated improvement



Hiring domain experts

It helps to have a principal domain expert

- Having a **directly-responsible individual (DRI)** helps you move faster
- They can build the best intuition for how your AI system performs
- Hire them as early as possible and give them ownership
- Your expert should do more than just look at data - they should help design and create your system

Hire more than “just” a domain expert

Your principal domain expert can also help with:

- hiring out a team of reviewers
- defining your sampling strategy for reviews
- analysing review data
- monitoring performance of reviewers
- steering product development
- prioritising eng work to improve AI performance
- talking to customers
- improving AI performance (through prompts, domain knowledge)

Hire more than “just” a domain expert

Your principal domain expert can also help with:

- hiring out a team of reviewers
- defining your sampling strategy for reviews
- analysing review data
- monitoring performance of reviewers
- steering product development
- prioritising eng work to improve AI performance
- talking to customers
- improving AI performance (through prompts, domain knowledge)

So it can be helpful if they have the following skills and experiences:

management/leadership

industry connections

statistics/data science

product skills/experience

communication skills

In Summary

- Most of the general eval principles apply to specialized verticals. There are added challenges with (i) defining quality and/or correctness, (ii) defining failure modes and (iii) writing prompts
- **Domain experts** perform a critical 'translation' step by looking at your **data** and converting it into actionable insights
- You can empower them through **custom review dashboards** which optimise for quality and speed of reviews and generate helpful **review data**
- That **review data** can prioritise work, facilitate AI iteration and even make automated improvements
- Hire a **principal domain expert** early, ideally with a breadth of skills beyond their domain expertise so they can drive the development of this system

Thank you

Dr Christopher Lovejoy, MD
Head of Clinical AI, Anterior

hi@chrislovejoy.me
www.chrislovejoy.me

